Title
Automated identification of pathways from quantitative genetic interaction data.

Journal
Molecular systems biology, 6(1)

ISSN
1744-4292

Authors
Battle, Alexis
Jonikas, Martin C
Walter, Peter
et al.

Publication Date
2010-06-01

DOI
10.1038/msb.2010.27

Peer reviewed

Molecular Systems Biology 6; Article number 379; doi:10.1038/msb.2010.27
Citation: *Molecular Systems Biology* 6:379

www.molecularsystemsbiology.com

molecular systems biology

Automated identification of pathways from quantitative genetic interaction data

Alexis Battle[1], Martin C Jonikas[2,3,4,5], Peter Walter[3,4], Jonathan S Weissman[2,4,5] and Daphne Koller[1,*]

[1] Department of Computer Science, Stanford University, Stanford, CA, USA, [2] Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA, USA, [3] Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA, [4] Howard Hughes Medical Institute, University of California, San Francisco, CA, USA and [5] California Institute for Quantitative Biomedical Research, San Francisco, CA, USA
* Corresponding author. Department of Computer Science, Stanford University, Stanford, CA 94305-9010, USA. Tel.: +1 650 723 6598; Fax: +1 650 725 1449; E-mail: koller@cs.stanford.edu

Received 7.9.09; accepted 7.4.10

High-throughput quantitative genetic interaction (GI) measurements provide detailed information regarding the structure of the underlying biological pathways by reporting on functional dependencies between genes. However, the analytical tools for fully exploiting such information lag behind the ability to collect these data. We present a novel Bayesian learning method that uses quantitative phenotypes of double knockout organisms to automatically reconstruct detailed pathway structures. We applied our method to a recent data set that measures GIs for endoplasmic reticulum (ER) genes, using the unfolded protein response as a quantitative phenotype. The results provided reconstructions of known functional pathways including N-linked glycosylation and ER-associated protein degradation. It also contained novel relationships, such as the placement of *SGT2* in the tail-anchored biogenesis pathway, a finding that we experimentally validated. Our approach should be readily applicable to the next generation of quantitative GI data sets, as assays become available for additional phenotypes and eventually higher-level organisms.
Molecular Systems Biology **6**: 379; published online 8 June 2010; doi:10.1038/msb.2010.27
Subject Categories: functional genomics; computational methods
Keywords: computational biology; genetic interaction; pathway reconstruction; probabilistic methods

Introduction

Recent developments have enabled large-scale quantitative measurement of genetic interactions (GIs) that report on the extent to which the activity of one gene is dependent on a second. It has long been recognized (Avery and Wasserman, 1992; Guarente, 1993; Phillips *et al*, 2000; Hartman *et al*, 2001; Segre *et al*, 2004; Tong *et al*, 2004; Drees *et al*, 2005; Schuldiner *et al*, 2005; Collins *et al*, 2007; St Onge *et al*, 2007; Jonikas *et al*, 2009; Costanzo *et al*, 2010) that functional dependencies revealed by GI data can provide rich information regarding underlying biological pathways. High-density GI maps systematically evaluate such interactions among a large set of genes. Further, the precise phenotypic measurements provided by quantitative GI data can provide evidence for even more detailed aspects of pathway structure, such as differentiating between full and partial dependence between two genes (Figure 1A) (Drees *et al*, 2005; Schuldiner *et al*, 2005; Collins *et al*, 2007; St Onge *et al*, 2007; Jonikas *et al*, 2009). As GI data sets become available for a range of quantitative phenotypes and organisms (Breslow *et al*, 2008; Roguev *et al*, 2008; Typas *et al*, 2008), such patterns will allow researchers to elucidate pathways important to a diverse set of biological processes. Methods based on RNAi will soon allow collection of similar data for human cell lines and other mammalian systems (Berns *et al*, 2004; Moffat *et al*, 2006; Firestein *et al*, 2008). Thus, computational methods for analyzing GI data could have an important function in mapping pathways involved in complex biological systems including human cells.

However, the tools for exploiting quantitative GI data have thus far not taken full advantage of the detailed information present in these measurements. The most commonly used approach is hierarchical agglomerative clustering (Tong *et al*, 2004; Schuldiner *et al*, 2005; Jonikas *et al*, 2009), which simply groups together genes whose interaction profiles are similar, and hence are likely to be involved in similar functions. More recent methods go beyond simple grouping to highlight aggravating or alleviating interactions between groups of related genes (Segre *et al*, 2004; Kelley and Ideker, 2005; Schuldiner *et al*, 2005; Qi *et al*, 2008). Although these forms of analysis produce a rough partition of genes into functional groups, they do not reveal the detailed structure of the pathways within the clusters or all the dependencies between

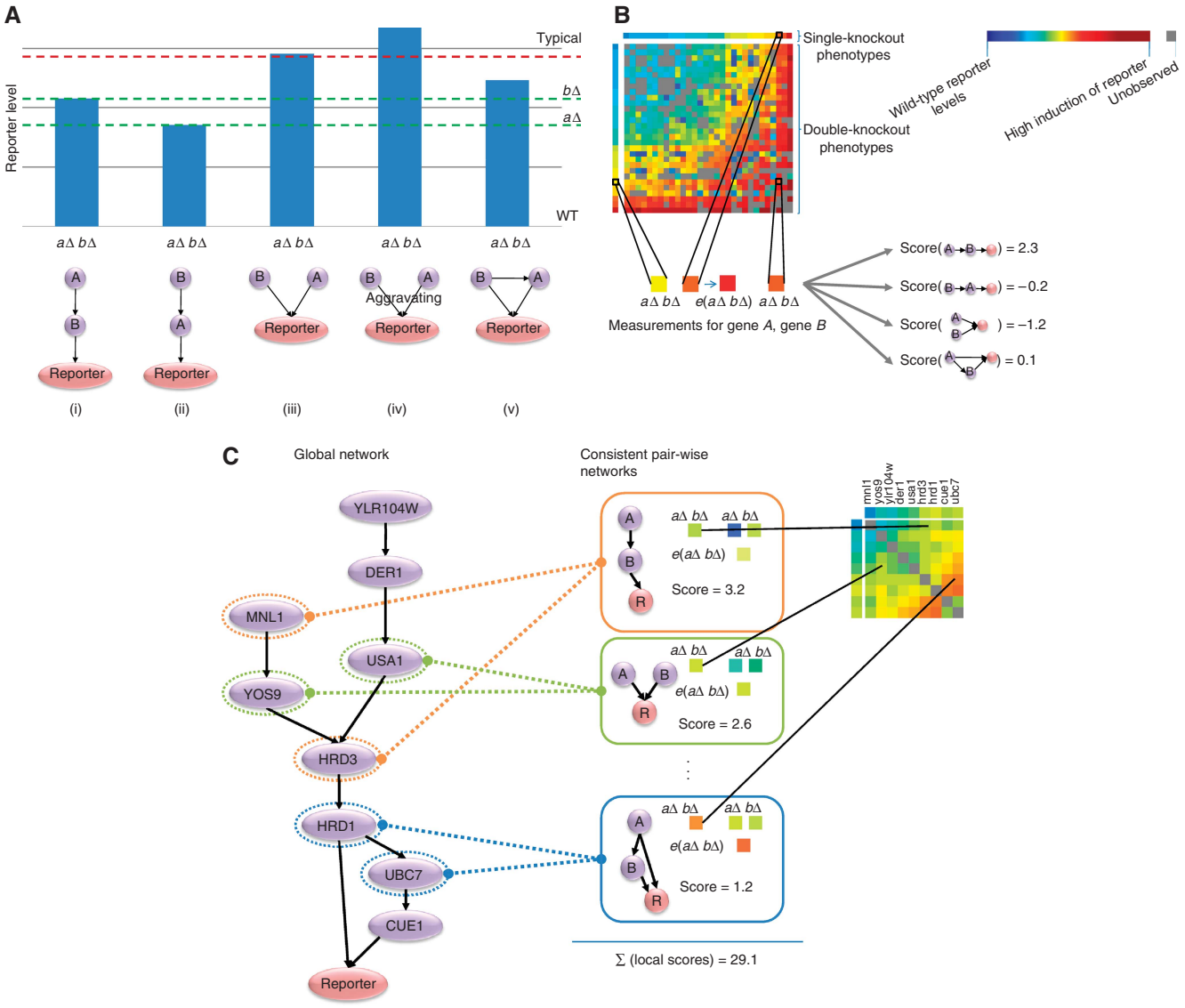


Figure 1 Overview of method. (**A**) Signature phenotypes for common pairwise relationships. Each pairwise relationship produces a 'signature' double knockout phenotype, as compared with observed individual knockout phenotypes (shown by dotted green lines) and the implied 'typical interaction' phenotype (dotted red line). (i, ii) Linear pathway configurations produce a double mutant phenotype similar to that of one of the single mutants. (iii) Independent actions result in a double knockout close to the expected (or 'typical interaction') phenotype. (iv) Genes acting separately but with related functions often result in aggravating interactions. (v) If the activity of one gene depends partially on the other (one gene also acts through a separate pathway), the double knockout is likely to be alleviating but not as fully as for a linear pathway. (**B**) Scoring pairwise structures with GI data. Using the double and single mutant measurements from a genetic interaction assay, a score is computed for each possible local graph structure for every pair of genes. For the example genes shown, the double knockout phenotype *a*Δ*b*Δ is very similar to the single *b*Δ. Thus, the linear pathway scores highly compared with the other possible pairwise structures. (**C**) Scoring complete activity pathway networks (APNs). Here, we show an APN over nine genes. Each complete APN is consistent with a set of local pairwise structures. For example, this graph is consistent with a pairwise relationship where *MNL1* is upstream of *HRD3* in a linear pathway. We evaluate the score of each consistent local relationship based on the corresponding two single and the double mutant reporter levels, and sum the local scores to compute the global score.

clusters. Other methods have used the more specific functional dependencies implied by quantitative GI data, but treat each pair of genes independently in inferring relationships (Drees *et al*, 2005; St Onge *et al*, 2007), not considering the consistency of each relationship with other data points or a global network model. Explorations of detailed multi-gene relationships have relied on manual examination, comparing observed GI values, or inferred pairwise relationships to a hypothesized multi-gene pathway model (Drees *et al*, 2005; St Onge *et al*, 2007; Jonikas *et al*, 2009); whereas manual analyses have shown considerable potential, they do not scale well, and require detailed *a priori* information. Another method, GenePath (Zupan *et al*, 2003), produces networks from small GI data sets, but does not automatically resolve the many conflicts that can arise from ambiguous and noisy evidence in large quantitative data sets. Thus, the above methods are not well suited for systematic automated reconstruction of pathways over large sets of genes. For further description, and more specific comparisons of related work to our method, see Discussion.

In this paper, we present a new method that exploits the high-quality, quantitative nature of recent GI assays (St Onge *et al*, 2007; Jonikas *et al*, 2009; Costanzo *et al*, 2010) to automatically reconstruct detailed multi-gene pathway structures, including the organization of a large set of genes into coherent pathways, the connectivity and ordering within each pathway, and the directionality of each relationship. We introduce activity pathway networks (APNs), which represent functional dependencies among a large set of genes in the form of a network. We present an automatic method to efficiently reconstruct APNs over large sets of genes based on quantitative GI measurements. This method handles uncertainty in the data arising from noise, missing measurements, and data points with ambiguous interpretations, by performing global reasoning that combines evidence from multiple data points. In addition, because some structure choices remain uncertain even when jointly considering all measurements, our method maintains multiple likely networks, and allows computation of confidence estimates over each structure choice. Thus, we can explore a range of structures consistent with our data, and focus on the highest confidence hypotheses for further investigation.

Results

The inputs to our method are the quantitative phenotype measurements over a set of single and double knockout organisms, as provided by a GI map. As described above, the APNs reconstructed by our method represent the functional dependencies among large sets of genes, and their combined effects on a downstream phenotype. We define an APN as a graph, with the activity of each gene corresponding to a node in the graph, and a special node representing the quantitative phenotype or *Reporter*. In an APN, a directed path between node *A* and node *B* represents a dependence of gene *A*'s activity on gene *B*'s activity, and if every path flowing from *A* to the Reporter passes through *B*, then gene *A*'s activity is fully dependent on *B*.

We now provide a basic outline of the APN reconstruction procedure (see Materials and methods for details). Overall, the method consists of first interpreting the GI data to derive a set of scores that represent preferences over the relationship between each pair of genes, and second searching for complete APNs that best satisfy these pairwise preferences. In the first phase, for every pair of genes we consider all possible pairwise network relationships (such as *B* follows *A* in a linear pathway), and compute a score statistically quantifying the extent to which their GI measurements support that relationship (Figure 1A and B). These statistical tests are based on the deviation of the observed double knockout phenotype from the outcome that would be expected for each network relationship (Figure 1A), according to the following assumptions. When two genes act in independent pathways, the effects of each mutation on the phenotype are compounded independently, frequently leading to a quantitative phenotype that is near a 'typical' level determined as a function of the phenotypes of the two individual mutants (Phillips *et al*, 2000; Collins *et al*, 2007; Jonikas *et al*, 2009) (Figure 1Aiii). Gene pairs that act separately but have related functions deviate substantially from such typical interactions, leading to so-called synthetic interactions, where the double mutant exhibits a more severe phenotype than expected (Guarente, 1993; Hartman *et al*, 2001; Tong *et al*, 2004) (Figure 1Aiv). Conversely, if the genes act in a single linear pathway, the effect of one gene is often mediated by the other gene, leading to an alleviating interaction where the double mutant displays a less dramatic phenotype than expected (Figure 1Ai, ii, v). In a subset of alleviating interactions, the double mutant has the same phenotype as one of the single mutants, indicating complete functional dependence (Avery and Wasserman, 1992; Segre *et al*, 2004; St Onge *et al*, 2007) (Figure 1Ai, ii). We note that if alternative interpretations of GI measurements were desired, our statistical tests could be adapted to accommodate them. Using similar tests, our method could also take advantage of measurements from more complex mutants, such as triple or quadruple knockouts, if they were available.

Given this initial computation of pairwise scores, we can now define a global score for full multi-gene APNs, and describe the procedure for finding high-scoring networks. We score a candidate APN *N* over the *full* set of genes by enumerating all pairwise network relationships that are consistent with *N*, and aggregating the corresponding pairwise scores into a global score (see Materials and methods, Figure 1C). Thus, the score of a network that encodes a certain relationship between two genes *A* and *B* will rely not only on the GI measurement for that pair, but on the constraints that relationship puts on all other relationships, and the corresponding GI measurements for those pairs. These other data points may strengthen support for weak or missing evidence from (*A*, *B*). For example, if a gene *C* strongly depends on both *A* and *B*, it could strengthen weak evidence for placing *A* in a linear chain with *B*. Conversely, if *C* depends strongly on *A*, but not on *B*, the combined evidence may reduce the support for placing *A* with *B*. In addition, as genes that are adjacent in the network interact in similar ways with other genes, we include a term that encourages genes with highly correlated GI profiles to be placed adjacent to each other in *N*. This term can also help compensate for missing or ambiguous pairwise scores. Despite the joint consideration of all evidence, in some instances, the global score may not provide conclusive support for a given structure, and in general there may be several APNs that are reasonably consistent with the data. Rather than select the single highest scoring APN, we sample the space of APNs using a Markov chain Monte Carlo (MCMC) method (Neal, 1998) (see Materials and methods), producing an ensemble of likelihood-weighted APNs from which we can infer confidence in any attribute, such as the presence of a particular linear pathway (Figure 2).

We applied our APN reconstruction method to the recent high-quality GI data set of Jonikas *et al* (2009), which examined the functional interaction between genes that contribute to protein folding in the endoplasmic reticulum (ER). Specifically, Jonikas *et al* used the cell's endogenous sensor (the unfolded protein response), to first identify several hundred yeast genes with functions in ER folding and then systematically characterized their functional interdependencies by measuring unfolded protein response levels in double

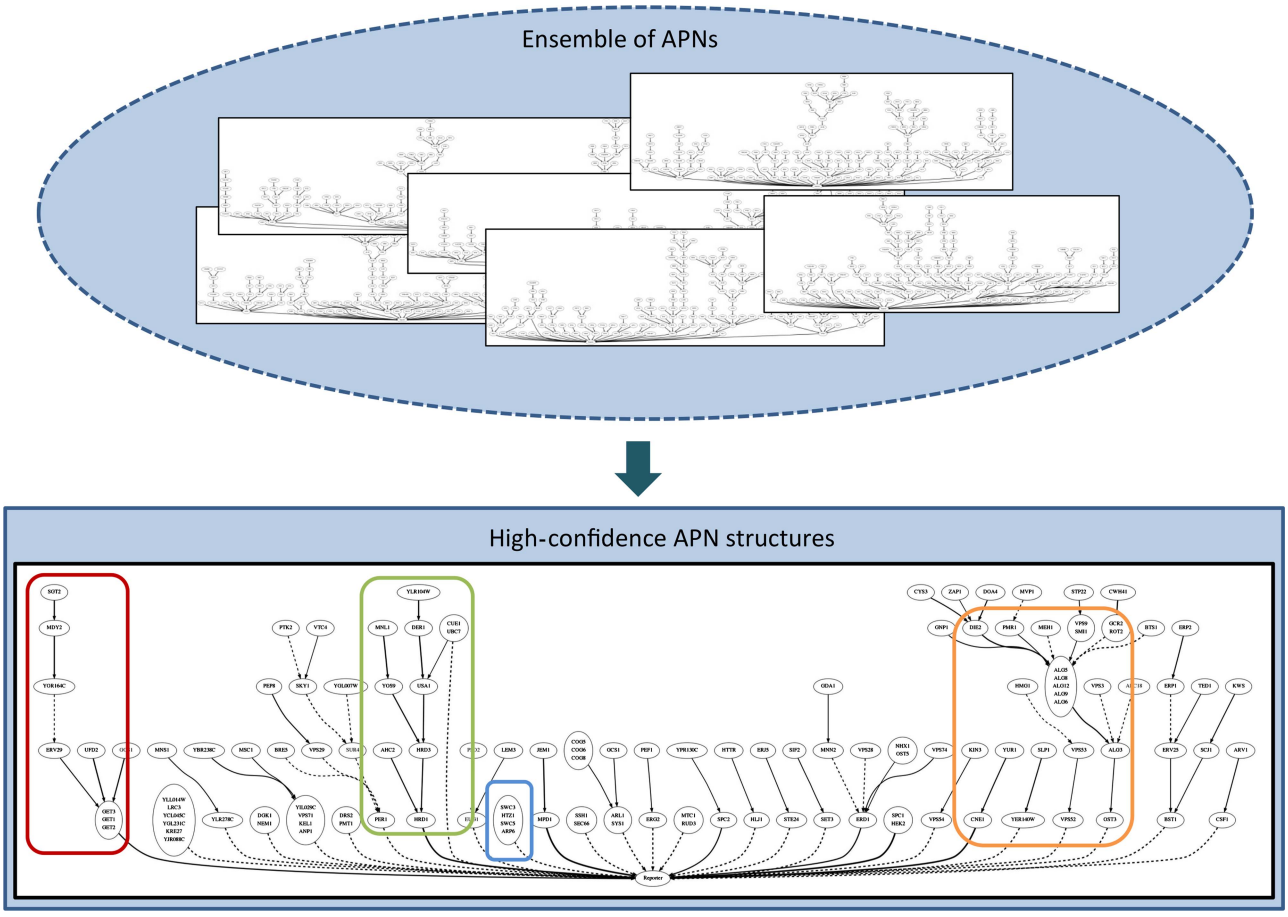


Figure 2 Activity pathway network ensemble for ER data. Applied to the data set of Jonikas *et al* (2009), our method produced an ensemble of 500 sampled APNs, each over 178 genes. Our method samples many full APNs from our probabilistic model, allowing us to estimate confidence over substructures. Using this likelihood-weighted ensemble, we produce confidence estimates for several graph substructures. For visualization, we produce an aggregated network, which highlights high-confidence pathways (see Materials and methods). Four interesting components of the high-confidence aggregated network have been highlighted, corresponding to pathways shown Figure 3—the blue box corresponds to Figure 3A, green to Figure 3B, orange to Figure 3C, and red to Figure 3D.

mutants (see Materials and methods). This analysis produced an ensemble of 500 likelihood-weighted APNs over 178 genes (Figure 2).

We evaluated the detailed structural predictions from our sampled APNs. The highest confidence structures (see Materials and methods) are visible in the aggregate network of Figure 2, and online (http://ai.stanford.edu/~ajbattle/APNgene_viz.html). Each of the relationships shown in Figure 3 (and discussed below) was automatically detected among the most likely subnetworks. By combining multiple weak, missing, or even contradictory measurements, our method does provide a global model that is much more robust than the individual measurements. For example, we predict full epistasis among several genes known to work together in the SWR complex, despite missing many of the relevant GI measurements (Figure 3A). Our APNs also identify a number of relationships that are not apparent from standard methods. As one example, our method places an edge between *NEM1* and *DGK1* with probability 0.61. Dgk1p phosphorylates diacylglycerol, an effect counteracted by the phosphatase Pah1p (Han *et al*, 2006), which is activated by Nem1p. Thus, in the absence of *DGK1*, *NEM1* has no function (reflected by their GI measurement). However, the GI profiles of *DGK1* and *NEM1* are only weakly correlated (0.05), so they are placed far apart in the clustering analysis of Jonikas *et al* (2009). In contrast, our algorithm highlights that *DGK1* fully masks *NEM1*. Such examples show the advantage of performing global reasoning, simultaneously considering evidence from all available measurements. Some relationships are evident from individual GI measurements, whereas others are supported by evidence from multiple data points or from correlation of GI profiles, and thus our method for reconstructing a global APN benefits from consideration of all evidence jointly.

We also performed an aggregate evaluation of our results by comparing to known biological relationships between gene pairs, including participation in pathways according to the Kyoto Encyclopedia of Genes and Genomes (KEGG), correlation of chemical genomic profiles in a recent high-throughput assay (Hillenmeyer *et al*, 2008), and similarity of Gene Ontology (GO) (Ashburner *et al*, 2000) annotations. Unfortunately, existing protein–protein interaction (PPI) data sets' (Gavin *et al*, 2006; Krogan *et al*, 2006) coverage of interactions among ER proteins is sparse and unreliable

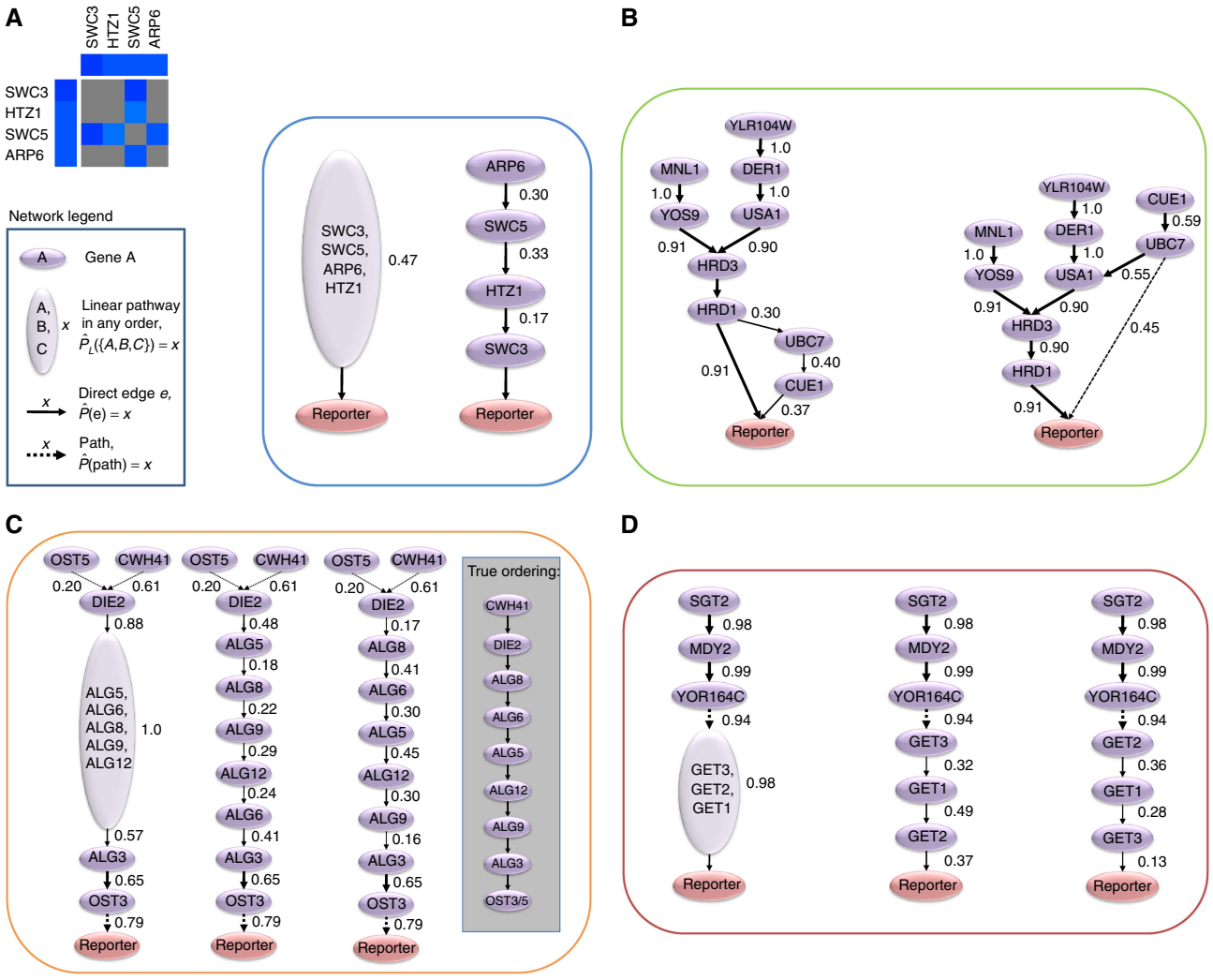
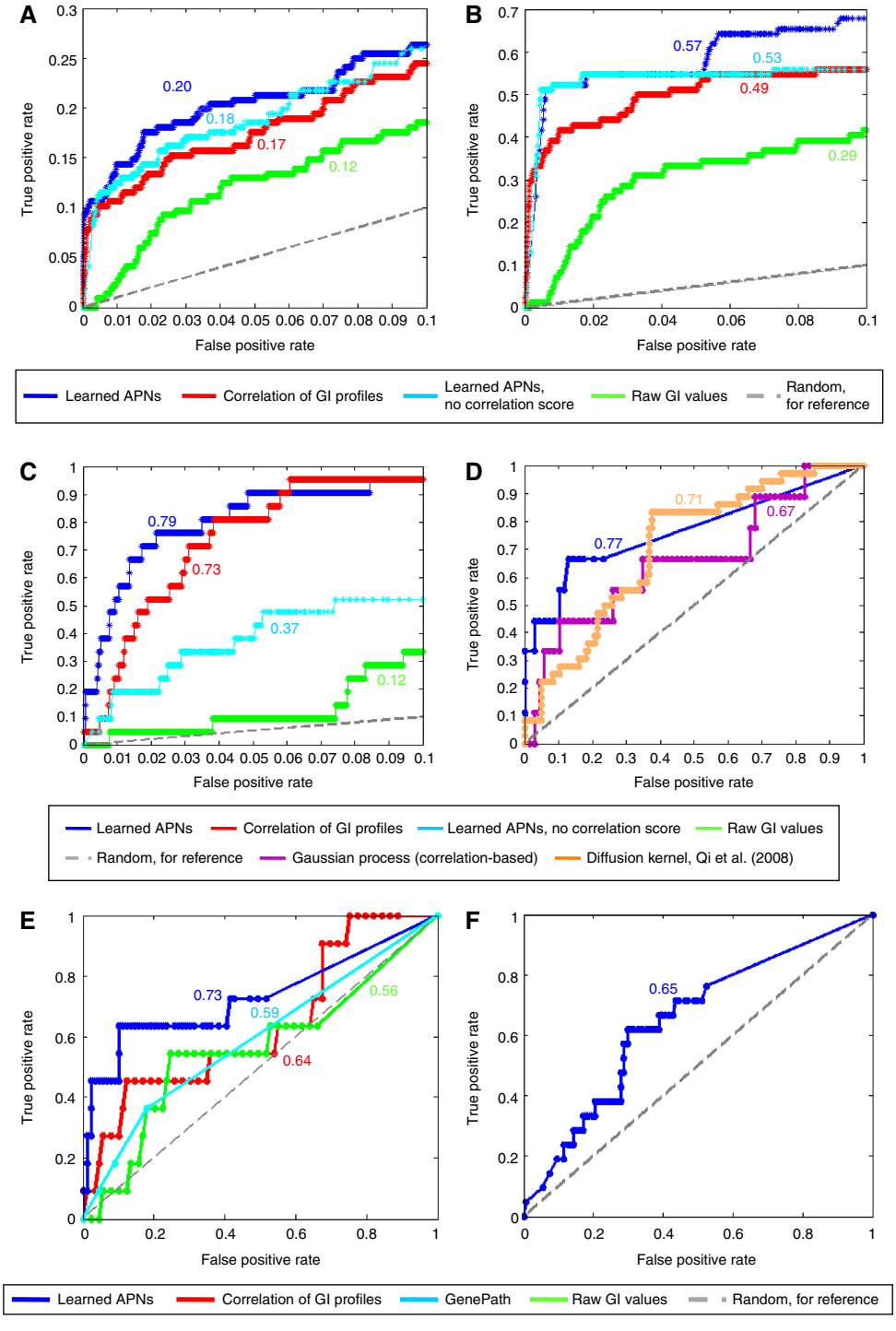


Figure 3 Reconstructed pathways for ER data. Visualization of reconstructed pathways. In each panel, we display the most likely network configurations for the relevant set of genes, according to our sampled APNs. A 'collapsed node' containing multiple gene names indicates a high-confidence linear pathway among the contained genes, but with the specific ordering varying among our samples. (**A**) SWR complex. APNs integrate data across multiple pairs of genes to discover relationships even if some data points are missing, statistically weak, or contradictory. Despite the unobserved combinations of *ARP6*, *SWC3*, and *HTZ1*, our method uses all available data, including the correlation scores and the observed alleviating interactions with *SWC5*, and places all four genes together in a linear chain, reflecting the known relationship among the SWR complex (which includes *SWC3*, *SWC5*, and *ARP6*) and the histone variant H2AZ (*HTZ1*). (**B**) ERAD pathway. Our reconstructed APNs placed several ERAD genes in common pathways with high confidence; we show the two most likely configurations of these pathways. Eight of these genes (*MNL1*, *YOS9*, *DER1*, *USA1*, *HRD1*, *HRD3*, *CUE1*, and *UBC7*) are known to be involved in ERAD function, and their respective placements in the graph are remarkably consistent with known interdependencies. The final gene, *YLR104W*, has also been suggested to participate in ERAD (Jonikas *et al*, 2009). (**C**) N-linked glycosylation pathway. Genes involved in N-linked glycosylation were automatically placed together in a single linear pathway with very high confidence, as shown in the aggregated view (left). The two highest probability detailed pathways (two middle networks) reflect many correct placements. The glucosyltransferase *DIE2* is robustly placed such that it is dependent on the other genes. *ALG9* and *ALG12* are correctly placed earlier, and *ALG3* is correctly placed at the start of this pathway with high confidence. *OST3* is correctly placed downstream, but *OST5* is incorrectly placed, likely because double mutant data with the other ALG genes was not available. For reference, the true ordering of this pathway (Helenius and Aebi, 2004) is shown as inset to the far right. (**D**) Tail-anchored protein insertion pathway. We show the three most likely configurations of the set. Very high confidence is assigned to the placement (and relative ordering) of *MDY2*, *YOR164c*, and *SGT2* upstream of *GET1*, *GET2*, and *GET3*. The relative ordering of *GET1*, *GET2*, and *GET3* is less certain, but they all occur in this linear pathway with probability 0.98 (leftmost network). *SGT2* is a poorly characterized gene not previously associated with tail-anchored protein insertion.

due to difficulty in isolation of membrane proteins, so we do not include an analysis of protein–protein interaction (PPI) data here. In each evaluation performed (Figure 4A–D), our reconstructed APNs were significantly more consistent with the known relationships than either the raw GI values or the Pearson correlation between profiles of GI values (commonly used in clustering analysis; Tong *et al*, 2004; Schuldiner *et al*, 2005; Jonikas *et al*, 2009). We also evaluated the COP score of Collins (Collins *et al*, 2007), but it gave nearly identical results to correlation (results not shown). To disambiguate the importance of the various components of our network score, we also compare our full method to a simplified version of our method excluding any correlation-based scoring. This does impact performance, but even the simplified version outperforms raw GI values on both data sets, and outperforms Pearson correlation on the GO and

KEGG analyses. Performance fell more drastically on the chemical profile analysis, perhaps unsurprisingly, as correlation of chemical profiles is strongly related to the similarity of genes' interactions. Next, we tested the ability of our procedure to predict unseen GIs. We constructed an ensemble of APNs using an initial set of GI data including 16 952 measurements, and used our APNs to predict the outcome of unseen GI measurements (see Materials and methods). We then performed additional GI experiments, and observed that our original reconstructed APNs made predictions that were much more consistent with the new GI values than predictions obtained by other approaches (see Materials and methods; Figure 4D). Finally, we note that sensitivity analysis suggests that our method is robust to noise beyond the level observed in the Jonikas *et al* (2009) data (Supplementary Figure 3).



Importantly, our approach provides not only an improved means for defining pairs or groups of related genes, but also enables the identification of detailed multi-gene network structures. In many cases, our method successfully reconstructs details of known cellular pathways, ranking them among the highest confidence structures (Figures 2 and 3). For example, it precisely reconstructs the known functional dependencies among components of the ER-associated degradation (ERAD) pathway (Nakatsukasa *et al*, 2008) (Figure 3B). *MNL1* is placed upstream of *YOS9*, consistent with existing data showing that *MNL1* generates the sugar species recognized by *YOS9* (Quan *et al*, 2008; Clerc *et al*, 2009). *YOS9*, *MNL1*, *DER1*, and *USA1* are placed upstream of *HRD3* and *HRD1*, consistent with data showing that degradation of certain substrates like CPY* requires all six components (Kim *et al*, 2005; Carvalho *et al*, 2006), whereas some substrates like Sec61-2 require only *HRD1* and *HRD3* but not *DER1*, *USA1* (Carvalho *et al*, 2006), or *YOS9* (Szathmary *et al*, 2005). The E2 Ubiquitin-conjugating enzyme *UBC7* and its membrane anchor *CUE1* are frequently placed downstream of the E3 Ubiquitin ligase *HRD1* and its regulatory partner *HRD3* in the learned APNs, consistent with the known function of Ubc7p and Cue1p in ubiquitination of Hrd1p substrates (Nakatsukasa *et al*, 2008). Finally, our algorithm recognizes that *HRD1* acts also through a distinct pathway not involving *UBC7* and *CUE1*, consistent with the ability of another E2 enzyme, Ubc1p, to partially substitute for *UBC7* and *CUE1* in their absence (Friedlander *et al*, 2000).

The pathway of biosynthesis of N-linked glycans (ALG) (Helenius and Aebi, 2004) presents a particular challenge for a pathway reconstruction algorithm. Although the different enzymes may act sequentially to assemble a glycan in the wild-type cell, the raw data of double mutant reporter levels suggests that some of the enzymes have residual activity in the absence of enzymes that act before them. Notably, the glucosyltransferases *ALG6*, *ALG8*, and *DIE2* appear to have some function in the *alg3*Δ, *alg9*Δ, and *alg12*Δ, possibly because of the branched nature of the high-mannose sugar that they generate. For example, the A branch remains intact in the *alg3*Δ, *alg9*Δ, and *alg12*Δ mutants (Burda and Aebi, 1999) and may still be partially glucosylated by Die2p in these strains, which could explain incomplete masking of Δ*die2* by these mutations. Nevertheless, our algorithm orders the ALG pathway remarkably accurately (Figure 3C), attesting to the robustness of our approach. We note that our algorithm also correctly detects a functional dependence of the *CWH41* Glucosidase I on all genes in the ALG pathway, despite the moderate correlation of *CWH41* with the ALG genes, which prevented it from being clustered with them (Jonikas *et al*, 2009). The relatively large number of genes involved in this pathway and the fact that it is well studied make this example amenable to quantitative analysis. We evaluated the ability of our method to predict each edge in the network (see Materials and methods) implied by the known ordering of this pathway (Helenius and Aebi, 2004), using direct links in the known pathway as positive edges, and treating all other possible links as negative edges. We compared predictions from our method with edges predicted from Pearson correlation alone and pairwise GI scores alone. We also applied GenePath (Zupan *et al*, 2003), using the subset of the data over the ALG genes, as the GenePath tool did not scale to our full data set. We computed ROC curves for each method (Figure 4E) and calculated the area under each curve (AUC), obtaining 0.7314 for our method, compared with 0.6399 for correlation, 0.5603 for GI scores and 0.5919 for GenePath, demonstrating that our method more accurately predicts the ordering and exact edges among the genes in this pathway.

More broadly, we analyzed the ability of our networks to predict details of known pathways, beyond the small well-studied glycosylation network discussed above. A more in-depth analysis of KEGG pathways indicates that our learned APNs are indicative of ordering in biological pathways. We compared all gene pairs (A, B), where A is found upstream of B in some KEGG pathway (the 'positive set'), to all gene pairs that are found together in some KEGG pathway, but where A is not upstream of B (the 'negative set'). Our learned APNs are much more likely to place A upstream of B for gene pairs in the positive set than pairs in the negative set (P-value 0.0218) (Figure 4F; Materials and methods).

Our results also suggest several novel relationships, including placement of uncharacterized genes into pathways, and novel relationships between characterized genes. These

Figure 4 Quantitative evaluation of learned APNs. For each ROC curve shown, the graph is annotated with the computed area under the curve (AUC). (**A**) Prediction of GO co-function. We evaluated the prediction of gene pairs, which share GO functional annotation. We compared prediction based on (1) the probability of placement of each gene pair in a shared pathway in the learned APNs, (2) Pearson correlation of GI profiles, (3) raw GI scores, and (4) placement in APNs learned without utilization of correlation scores. We restricted AUC computations to the false-positive range shown, obtaining normalized areas 0.202, 0.173, 0.117, and 0.182, respectively. (**B**) Prediction of KEGG pathway membership. We evaluated the prediction of gene pairs, which participate together in some KEGG canonical pathway. We compared prediction based on (1) the probability of placement of each gene pair in a shared pathway in the learned APNs, (2) Pearson correlation of GI profiles, (3) raw GI scores, and (4) placement in APNs learned without utilization of correlation scores. We restricted area under the curve (AUC) computations to the false-positive range shown, obtaining 0.572, 0.494, 0.292, and 0.529, respectively. (**C**) Prediction of similar chemical sensitivity phenotypes. On the basis of the data set of Hillenmeyer *et al* (2008, we selected pairs of genes with highly similar chemical phenotypes. We compared the ability of four methods to predict membership in this test set—probability of placement in a shared pathway in the learned APNs, Pearson correlation from GI profiles, raw GI scores, and placement in APNs learned without correlation scoring. The normalized AUCs for the displayed range were 0.792 (APN), 0.725 (correlation), 0.118 (GI), and 0.371 (APN without correlation). (**D**) Prediction of unknown genetic interactions. For a set of measurements unavailable at the time of APN learning, we compared methods for predicting unseen alleviating interactions. We compare ROC curves for predictions made from (1) learned APNs, where we score each pair of nodes according to the probability of placement in a shared pathway according to the APNs; (2) predicted GI values from Gaussian Process regression (Williams and Rasmussen, 1996), a baseline method that uses the correlation of *observed* GI profiles; and (3) predicted interactions based on the diffusion kernel method (Qi *et al*, 2008). The resulting AUCs were 0.77, 0.67, and 0.71, respectively. (**E**) Prediction of N-linked glycosylation pathway edges. We evaluated the prediction of edges in the N-linked glycosylation pathway (Helenius and Aebi, 2004). We compared prediction based on (1) the probability of an edge between each gene pair in the learned APNs, (2) Pearson correlation of GI profiles, (3) raw GI scores, and (4) GenePath predictions (Zupan *et al*, 2003). We obtained AUCs of 0.7314, 0.6399, 0.5603, and 0.5919, respectively. (**F**) Prediction of KEGG pathway ordering. We evaluated the ability of our networks to predict ordering *within* KEGG pathways, and obtained an AUC of 0.6480. Our results are significant with P=0.0218.

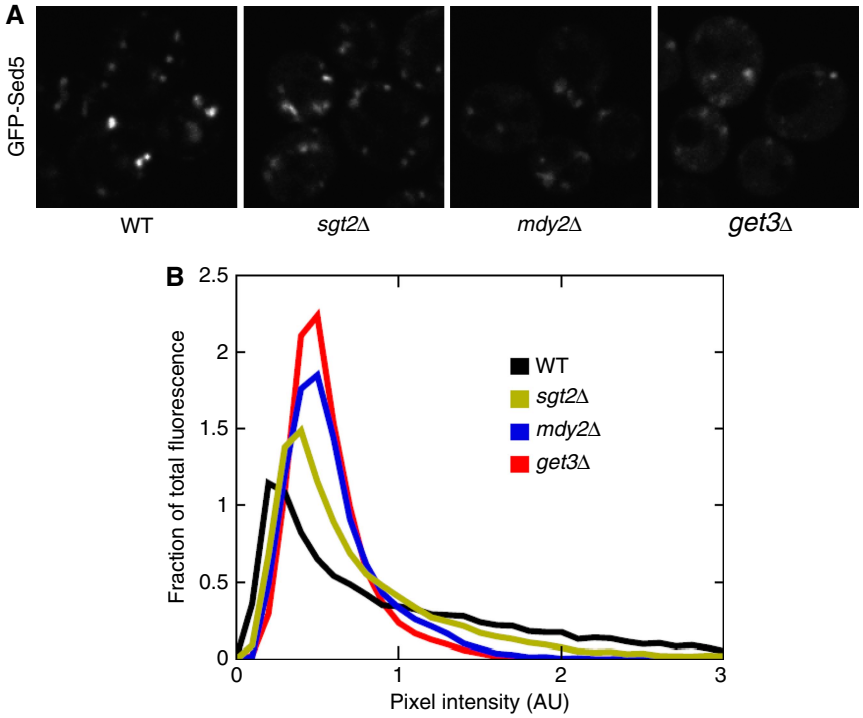


Figure 5 GFP-Sed5p localization defect in *sgt2*Δ. (**A**) Microscopy. GFP-Sed5p localization in WT, *sgt2*Δ, *mdy2*Δ, and *get3*Δ strains demonstrating a defect in GFP-Sed5p localization in *sgt2*Δ. These results support the placement of *SGT2* in the tail-anchored protein biogenesis pathway shown in Figure 3D. (**B**) Quantitative analysis. The images of at least 30 cells per strain with similar average fluorescence were quantified to determine the distribution of each strain's total fluorescence across pixels of different intensities. The distribution of fluorescence in the *sgt2*Δ strain differs from that of the wild-type strain with $P < 1e{-}13$, and is similar to the distribution for the knockout strains of other genes known to be involved in this pathway.

include the dependence of the J domain chaperone *JEM1* on the PDI homolog *MPD1*, dependence of the Ubiquitin-recycling enzyme *DOA4* on N-linked glycosylation, and the dependence of the E3 Ubiquitin ligase *DOA10* on the signal peptidase complex subunit *SPC2*. Our APNs also place the poorly characterized TPR-containing protein *SGT2* upstream of the tail-anchored protein biogenesis machinery components *GET3*, *GET4*, and *MDY2* (also known as *GET5*) (Stefanovic and Hegde, 2007; Schuldiner *et al*, 2008; Jonikas *et al*, 2009) (Figure 3D), suggesting that *SGT2* has a function in the insertion of tail-anchored proteins into membranes. Consistent with this prediction, our experimental analysis shows that *sgt2*Δ cells show a defect in the localization of the tail-anchored protein GFP-Sed5 from punctuate Golgi structures to a more diffuse pattern, as seen in other genes involved in this pathway (Figure 5; Supplementary Figure 1). Although this manuscript was under review, Costanzo *et al* (2010) have independently suggested participation of *SGT2* in this pathway. Their results also show a direct physical interaction between *SGT2* and *MDY2*, further supporting the precise placement of *SGT2* in our networks.

Discussion

Our results show that multi-gene, detailed pathway networks can be reconstructed from quantitative GI data, providing a concrete computational manifestation to intuitions that have traditionally accompanied the manual interpretation of such data. In particular, our automatically reconstructed APNs for the ER recapitulate the details of several important ER pathways and show the ability of our analysis to produce testable biological hypotheses. Our reconstructed networks are available for visualization and analysis online (http://ai.stanford.edu/~ajbattle/APNgene_viz.html).

Our APN method builds on results from several important pieces of work (Avery and Wasserman, 1992; Guarente, 1993; Phillips *et al*, 2000; Hartman *et al*, 2001; Zupan *et al*, 2003; Segre *et al*, 2004; Tong *et al*, 2004; Drees *et al*, 2005; Schuldiner *et al*, 2005; Collins *et al*, 2007; St Onge *et al*, 2007; Jonikas *et al*, 2009), some of which are particularly relevant to our method. For example, Drees *et al* (2005) and St Onge *et al* (2007) provide frameworks that, similar to the first stage of our method, examine individual GI measurements to identify likely functional relationships for each pair of genes, including some subtypes of alleviating interactions. In a separate computation, Drees *et al* (2005) also quantify the similarity between pairs of GI profiles. Collins *et al* (2007) explore the use of both correlation and individual GI measurements in identifying pairs of co-functional genes. Each of these studies was able to identify biologically relevant patterns, often highly predictive of gene function. The GenePath work (Zupan *et al*, 2003) also showed that network structures can be derived from GI measurements, though on a small scale (fewer than 20 genes). All of the above results provided evidence that the individual GI measurements and the similarity between GI profiles can both provide detailed information regarding underlying relationships between genes. However, these

methods do not automatically consider *all* evidence jointly, disambiguating borderline cases and constructing a global model of the effects of all genes on the measured phenotype. Our method, on the other hand, does perform such global reasoning in a statistically robust framework, and thus our APNs reveal details not expressed by the other models, such as ordering within a multi-gene linear pathway.

One limitation of our analysis is that the relationships represented by the edges in the APNs may sometimes be difficult to interpret, as they may not correspond directly to specific physical relationships. This limitation reflects the fact that GIs themselves can arise from a broad range of functional relationships, including sequential interaction with a common substrate (as in N-linked glycosylation), direct PPI (as in the SWR complex), or indirect dependencies (as in the *NEM1*, *DGK1* example). In addition, because the ordering of genes in an APN corresponds to *dependency*, a pathway in an APN may sometimes reverse the ordering relative to the order of action in the underlying biological process, depending on the mechanisms involved (Avery and Wasserman, 1992). Although networks representing specific physical relationships have been reconstructed from other types of data (Beyer *et al*, 2006), GI measurements provide a more direct indication of the *functional* dependencies between genes. Other work (Kelley and Ideker, 2005) has successfully combined GI measurements with PPI information. Similarly incorporating PPI and other data into our method could likewise provide more specific interpretations for some of the relationships encoded in our detailed APNs, and is a direction for further investigation.

Ongoing technological developments in both genetics and imaging are enabling the measurement of GI data at a genome-wide scale, using high-accuracy quantitative phenotypes that relate to particular biological functions. These methods can help elucidate a broad range of pathways by using different molecular phenotypes as reporters. Using new methods such as RNAi-based approaches, such assays will soon be available for higher-level organisms, including human cells (Berns *et al*, 2004; Moffat *et al*, 2006; Firestein *et al*, 2008). A method that provides a high-quality *de novo* reconstruction of functional gene networks can thus provide an important tool in understanding human pathways.

Materials and methods

Data set

The data set of Jonikas *et al* (2009) contains 444 genes that were observed to significantly change the UPR phenotype. For these genes, the phenotypes of 86 396 double mutants were measured. Jonikas *et al* also compute, for each double mutant $a\Delta b\Delta$ the *typical* GI value $e(a\Delta b\Delta)$. This typical value represents the case where the two genes do not interact genetically. Roughly, the function is computed using a standard multiplicative model of the reporter levels for the two individual mutants, modified by incorporating a Hill function to model the saturation of the reporter signal. We use the same function and parameters fit to the overall GI data set by Jonikas *et al* (2009). The majority of pairs lie close to the typical interaction level, leaving only a small set of pairs that deviate significantly, thus being plausible candidates for GIs. Of the 444 genes in the data set, we selected those that induced UPR beyond wild-type levels, displayed an alleviating interactions with *HAC1* or *IRE1* (indicating dependence on the transcriptional regulators of the reporter), were measured against at least 40 other genes, and whose data fell close to their typical interaction curves. In addition, we threw out all data for gene pairs where the cell count was low (fewer than 350 per well) in over 60% of the attempted measurements. The resulting data set comprised 178 genes and 20 778 GI measurements.

Activity pathway network

We represent an APN as a Bayesian network (BN) (Pearl, 1988). For each gene in our data set, a random variable capturing the activity of its gene product is represented by a node in the network. In addition, we include one node (*Reporter*) representing the measured quantitative phenotype (in our case the UPR reporter). Because of the sparsity of the measurements—each experiment provides activity levels only for the reporter and the two deleted genes—we do not attempt to reconstruct the parameters quantifying the interactions between the variables of the network. Standard BN learning methods that handle incomplete data (Friedman, 1998) are unsuitable when most of the values are unobserved, and, indeed, performed very poorly when applied to our GI data (results not shown). Rather, we use the conditional independence assumptions encoding the *structure* of the BN (Pearl, 1988; Sprites *et al*, 1993), seeking to find a network that encodes independence assumptions that are well supported by the GI measurements. For instance, if gene A appears fully epistatic to gene B in the data, the network should indicate that the reporter level is independent of the activity level of B given the activity level of A, an independence property that is encoded by a linear pathway structure. The '*Reporter*' node is always a descendant of all other nodes in our network, as GI measurements reflect only downstream effects of other genes on the reporter.

Scoring of pairwise network structures

The score of each candidate APN N is derived from local scores over pairwise relationships in the network structure, based on data for each corresponding pair of genes. For a given pair of genes A and B, we consider all possible relationships r (listed in Table I below), which may characterize their relationship in network N. For each such pairwise relationship, we defined a quantitative reporter value $\mu_r(a\Delta b\Delta)$ that we expect for a gene pair A, B with this relationship. The expected values are defined in terms of the reporter values of the two individual mutants, denoted $R(a\Delta)$ and $R(b\Delta)$, and the typical interaction value $e(a\Delta b\Delta)$, which is computed as in Jonikas *et al* (2009). In detail, $\mu_r(a\Delta b\Delta)$ is defined according to Table I.

Note that these expected values μ_r are in agreement with the conditional independencies encoded by the BN structure. For example, if A and B are in a linear pathway with A downstream, the conditional independencies encoded by the BN imply that the reporter is independent of B given A. This implies that in the context of knocking out gene A (which fixes its activity to zero), altering the activity of gene B has no further affect on the reporter, and thus $\mu_r(a\Delta b\Delta)=R(a\Delta)$.

In each case, the score given to relationship r for pair (A, B) is based on a Gaussian distribution, where we evaluate the probability of observing the actual measured reporter value $R(a\Delta b\Delta)$ given the signature outcome $\mu_r(a\Delta b\Delta)$. Specifically, we evaluate

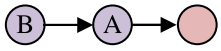
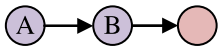
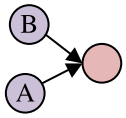
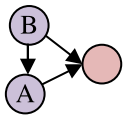
$$\text{score}(r) = -\frac{1}{2\sigma_r^2}\left(R(\Delta a\Delta b) - \mu_r(\Delta a\Delta b)\right)^2 + \log(p(r)).$$

The variance σ_r^2 used was the empirical variance of repeated reporter level measurements, and $\log p(r)$ represents the prior probability of any given relationship. The distribution $p(r)$ was estimated directly from frequencies obtained from the data by classifying each data point according to the relationship whose mean fell closest to the observed GI measurement. In the case of missing pairwise GI measurements, the corresponding score will simply be $\log p(r)$.

Global APN score

We constructed a distribution over complete APNs N based on the local scores described above. More precisely, for candidate APN N, we find all consistent pairwise network structures—that is, for each pair of

Table I Expected reporter levels

Relationship r	Network structure	$\mu_r(a\Delta b\Delta)$
A and B in a linear pathway, A downstream		$R(a\Delta)$
A and B in a linear pathway, B downstream		$R(b\Delta)$
A and B affect the reporter separately		$e(a\Delta b\Delta)$
A and B are partially interdependent, but each also has a path to the reporter not dependent on the other		$0.5 \times (e(a\Delta b\Delta) + \max(R(a\Delta), R(b\Delta)))$

For each possible relationship r between a pair of genes, we show the corresponding network structure visually, and specify the expected reporter level $\mu_r(a\Delta b\Delta)$. For a candidate APN, these values are used to compute local scores for the pairwise relationships implied by the APN, and the local scores are then combined to compute the global APN score.

genes, we determine which pairwise relationship r holds in N (linear pathway, independent, or partial dependence), and compute the pairwise score, score(r). Note that to determine which relationship holds between genes A and B in N, we must consider all paths from A to '*Reporter*' node and from B to '*Reporter*,' which may run through several other nodes, as shown in Figure 1C. In addition to score(r), we compute an additional score for every possible edge e in N, based on the observation that if two nodes are directly linked in N, their structural relationships to other nodes in the network will be similar, and thus on expectation their GI measurements will also be similar. To explain this intuition at a high level, if N includes an edge from A to B, then the set of nodes that depend on A and the set that depend on B will overlap significantly. For instance, every ancestor of A must be an ancestor of B, and every descendant of B must be a descendant of A. This implies that we expect gene A and gene B to have significant GIs with similar sets of genes, and thus that their GI profiles are likely to be similar. To encode a structural preference for placing an edge between nodes with similar GI profiles, for each edge e in N, we compute score(e) based on the Pearson correlation between the GI profiles of the two genes connected by e. We also penalize score(e) by a constant (C=2.8) to penalize overly complex or dense networks.

The final distribution is defined as

$$P(N) = \frac{1}{Z} \exp\left(\sum_{r \text{ consistent with } N} \text{score}(r) + \sum_{e \in \text{ edges}(N)} \text{score}(e) \right).$$

The use of this correlation-based score(e) in addition to score(r) helps handle missing data and borderline measurements (where multiple relationships r are similarly plausible). One could imagine adjusting the balance between the strength of score(e) and score(r) in our method, but due to a lack of supervised data, we simply left them equally weighted to avoid excessive manual tuning.

Sampling

Given the distribution $P(N)$ defined above, we apply annealed importance sampling (AIS) (Neal, 1998) to collect fully specified APN structures, which induces an exponentially large state space. In addition, $P(N)$ defined above is likely to have many modes; thus, AIS is particularly appropriate for our K APNs sampled randomly from $P(N)$. AIS is an MCMC method designed to produce independent samples even when applied to complex multi-modal distributions. By using multiple independent annealing runs, AIS helps address the slow mixing time often encountered in using MCMC for such distributions.

To sample from a desired distribution such as $P(N)$, AIS uses a sequence of helper distributions $q_1(N) \ldots q_p(N)$, and Markov chain transition probabilities T_j for which the corresponding q_j is invariant. To generate a single importance weighted sample, the AIS procedure specifies that we sample an initial network n_0 directly from q_p, and then sequentially apply transitions according to $T_{p-1} \ldots T_1$. We compute an importance weight w as we go, and take the network n_1 resulting from the final transition from T_1 as our sample. Although in principle we can use arbitrary q_j, the procedure provides good estimates if q are increasingly good estimates of $P(N)$. Thus, we used the distributions q_p=uniform, $q_0 \propto P(N)$, and $q_j = q_0^{\beta_j} q_p^{1-\beta_j}$ for a sequence $1=\beta_0 > \beta_1 > \ldots > \beta_p=0$. We used settings p=1000, and β_j falling off exponentially according to $\exp(-j/200)$.

We now describe the construction of T_j for our domain. At each step $j<p$, T_j is specified using a standard Metropolis–Hastings construction. We propose a modification n' to the existing structure n_j, according to a distribution $S(n, n')$. The modification is accepted, and n_{j-1} is set to n', with probability

$$\frac{q_j(n_{j-1}) S(n_{j-1}, n_j)}{q_j(n_j) S(n_j, n_{j-1})}$$

Otherwise, the move is rejected and $n_{j-1}=n_j$.

Our proposal distribution $S(n, n')$ remains fixed for all steps j. Inspired by the particular form of our structure score, we included some non-standard structure search operators to generate a candidate n' from n. We considered the following operators: inserting a node into a linear pathway, popping a node out of a linear pathway, swapping the order of two nodes in a linear pathway, detaching an entire linear pathway and reattaching it elsewhere in the network, and adding or deleting an edge. At each step in the sampling procedure then, our proposal distribution S is the uniform distribution over legal networks that can be constructed by applying any single structure search operator to n_j.

At the end of the annealing schedule of each independent run k, we sample a single APN $N_k=n_1$, and save the final importance weight as w_k. The confidence in any given substructure g is then computed as

$$P(\text{substructure } g) = \frac{\sum_{k=1}^{K} w_k \cdot \delta[N_k \text{ consistent with } g]}{\sum_{k=1}^{K} w_k}$$

where δ is the indicator function. In many cases, including evaluations described in this paper, we are interested in the probability that a set of genes **G** occur together in a single linear pathway (in any order). We will denote this $P_L(G)$, which we find by computing

$$\hat{P}_L(G) = \frac{\sum_{k=1}^{K} w_k \cdot \delta[\mathbf{G} \text{ occur in single linear chain in } N_k]}{\sum_{k=1}^{K} w_k}.$$

For the ER data set, we sampled K=500 APNs. Computational analysis shows that our AIS sampler produces highly repeatable estimates of the probabilities of different structural features (Supplementary

Figure 2), indicating that our runs are sampling from the specified posterior distribution.

Pseudocode for APN reconstruction

Pre-compute score(r) for every pair of genes, and each possible relationship r

Pre-compute score(e) for every pair of genes
AIS procedure:
$P(N)$ defined by pre-computed score(r), score(e)
Specify distributions q_{p-1}.. q_1 to approach $P(N)$
For k=1 … K:
p:=1000
Sample initial network n_p from distribution q_p=uniform
For j=p−1 … 1:
Propose network N_j according to uniform distribution over legal structure moves from N_{j-1}
Accept N_j according to Metropolis–Hastings equation for q_j
Update importance weight w
Save sample N_k= n_1 and w_k
Return samples networks N_1… N_K and importance weights

Computation of likely pathways and confidence estimates

See Supplementary information for MATLAB implementation.

Missing GI value prediction

Using an initially available set of GI measurements (a subset of the data for our full experiments), with 16 952 measurements observed, we applied our method to generate an ensemble of 500 APNs. We then evaluated, for each pair of genes (A, B), the probability $\hat{P}_L(\{A,B\})$ that they are in a linear pathway. For comparison, we implemented Gaussian process (Williams and Rasmussen, 1996) regression using a kernel constructed from the Pearson correlation of GI profiles, and used this to predict unseen GI values. We then performed the additional GI assays, and constructed a test set of 78 measurements. We selected the test gene pairs A, B such that the original data set contained several measurements for both genes—that is for many genes C, we observed $R(a\Delta c\Delta)$ and for many D we observed $R(b\Delta d\Delta)$. We required that the geometric mean of the number of these training measurements for A and the number of training measurements for B by at least 180. Within this set, we identified a positive set of gene pairs with significant alleviating GIs, defining an interaction to be alleviating if the observed GI interaction score was negative with a magnitude greater than $|R(\Delta a\Delta b)-\max(R(\Delta a), R(\Delta b))|$. We then attempted to predict whether a test interaction was alleviating using the probability $\hat{P}_L(\{A,B\})$ derived from our APNs, and produced an ROC curve. We compare these results with those obtained using Gaussian processes (Williams and Rasmussen, 1996) as described above, and with results from the diffusion kernel GI prediction method of (Qi *et al*, 2008).

GO co-functionality evaluation

Using GO biological process annotations (Ashburner *et al*, 2000), we identified all gene pairs that occur in some GO functional group of 20 genes or less, using these pairs as positive examples and all other pairs as negative. From our learned networks, we computed $\hat{P}_L(\{A,B\})$ for every pair of genes, and evaluated this score as a predictor of GO co-functionality, generating an ROC curve. As shown in Figure 3A, we also evaluated the raw GI score of each pair and the correlation of GI profiles. Finally, we evaluated networks learned without the use of the correlation component $P(N)$, score(e).

KEGG pathway analysis

We used canonical pathways from the KEGG (Kanehisa and Goto, 2000) to validate our learned networks. First, we identified all gene pairs that occur together in some KEGG pathway, and used these pairs as positive examples and all others as negative. We then computed ROC curves (Figure 3B) as in the GO analysis. In addition, we evaluated the ability of our networks to predict ordering *within* pathways (Figure 3F). KEGG provided 21 gene pairs (A, B) that occur in a pathway together and where gene A is known to be upstream of gene B, and 147 gene pairs that occur in a pathway together but gene A is not known to be upstream of gene B (see Supplementary information). We compared the distribution of P (A upstream of B), according to the sampled networks, for these two sets. We found that the median value for the positive set was 0.35, and for the negative set was 0.004, and the two distributions were found to be different according to the Whitney–Mann test with P=0.0218.

Chemical phenotype evaluation

Using the data set of Hillenmeyer *et al* (2008), we identified all gene pairs whose chemical phenotype correlation was reported as $\geqslant 0.7$, using these pairs as positive examples and all others as negative. ROC curves were computed as for GO co-functionality prediction.

Post-processing and visualization

For confidence estimates presented in this manuscript, we identified interesting substructures and compute $\hat{P}$(substructure g) as described above. To visualize our results (Figure 2; http://ai.stanford.edu/~ajbattle/APNgene_viz.html), we also clustered APNs using hierarchical agglomerative clustering, with features of each sampled APN composed of the list of edges and the list of separating relationships present in the APN. Edges and paths shown all occur with $P>0.5$ according to the aggregated samples. For display, we also collapse sets of nodes $\mathbf{G}$ where both $\hat{P}_L(\mathbf{G})>0.6$, indicating that the genes in $\mathbf{G}$ are likely to form a linear pathway, and $\hat{P}_L(\mathbf{G})>1.8\times P$(specific ordering of $\mathbf{G}$), indicating that no particular ordering within that pathway is dominant. These two thresholds were chosen by hand, simply to provide empirically clean visualizations, and do not affect any other estimates reported.

To find structures of interest (Figure 3), we extracted the linear chains and edges found with highest confidence P(substructure) among the sampled networks. In general, although it is useful to visually examine high-scoring networks, we note that for any specific network relationship of interest, it is essential to use the entire ensemble of APNs to estimate P(substructure) as a measure of confidence. Unless this estimate is high, the presence of that relationship in a single high-scoring APN may be a fluke, and may not even contribute to the high score of that APN.

Experimental procedures

SGT2, *MDY2*, and *GET3* were deleted in strain BY4741 (WT) (Brachmann *et al*, 1998) using the nourseothricin marker using standard methods (Goldstein and McCusker, 1999). Plasmid pMS113, containing HA-GFP-Sed5 on a pRS315 backbone, was transformed into these strains. Strains were grown in SD–LEU. Imaging and quantification were performed as described earlier (Jonikas *et al*, 2009).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank V Jojic for useful discussions, as well as the members of the Koller, Weissman, and Walter research groups. This work was supported by the NSF (DK, AB, and MCJ), the HHMI (JSW), and specifically NSF grant DBI-0345474 (DK and AB).

Author contributions: AB, MCJ, DK, and JSW designed analysis approach, evaluated results, and wrote the paper. AB and DK developed statistical methods. AB wrote analysis code. MCJ performed

GFP-Sed5 experiment. MCJ, JSW and PW designed and collected the GI data set.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* **8:** 312–316

Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, Agami R, Ge W, Cavet G, Linsley PS, Beijersbergen RL, Bernards R (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428:** 431–437

Beyer A, Workman C, Hollunder J, Radke D, Möller U, Wilhelm T, Ideker T (2006) Integrated assessment and prediction of transcription factor binding. *PLoS Comput Biol* **2:** e70

Brachmann C, Davies A, Cost G, Caputo E, Li J, Hieter P, Boeke J (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14:** 115–132

Breslow D, Cameron D, Collins S, Schuldiner M, Stewart-Ornstein J, Newman H, Braun S, Madhani H, Krogan N, Weissman J (2008) A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods* **5:** 711–718

Burda P, Aebi M (1999) The dolichol pathway of N-linked glycosylation. *Biochimica et Biophysica Acta* **1426:** 239–257

Carvalho P, Goder V, Rapoport T (2006) Distinct ubiquitin-ligase complexes define convergent pathways for the degradation of ER proteins. *Cell* **126:** 361–373

Clerc S, Hirsch C, Oggier D, Deprez P, Jakob C, Sommer T, Aebi M (2009) Htm1 protein generates the N-glycan signal for glycoprotein degradation in the endoplasmic reticulum. *J Cell Biol* **184:** 159–172

Collins S, Miller K, Maas N, Roguev A, Fillingham J, Chu C, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger S, Hieter P, Zhang Z *et al* (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446:** 806–810

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear E, Sevier C, Ding H, Koh J, Toufighi K, Mostafavi S, Prinz J, St Onge R, VanderSluis B, Makhnevych T, Vizeacoumar F, Alizadeh S, Bahr S, Brost R, Chen Y, Cokol M *et al* (2010) The genetic landscape of a cell. *Science* **327:** 425–431

Drees BL, Thorsson V, Carter GW, Rives AW, Raymond MZ, Avila-Campillo I, Shannon P, Galitski T (2005) Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol* **6:** R38

Firestein R, Bass A, Kim S, Dunn I, Silver S, Guney I, Freed E, Ligon A, Vena N, Ogino S, Chheda M, Tamayo P, Finn S, Shrestha Y, Boehm J, Jain S, Bojarski E, Mermel C, Barretina J, Chan J *et al* (2008) CDK8 is a colorectal cancer oncogene that regulates [bgr]-catenin activity. *Nature* **455:** 547–551

Friedlander R, Jarosch E, Urban J, Volkwein C, Sommer T (2000) A regulatory link between ER-associated protein degradation and the unfolded-protein response. *Nat Cell Biol* **2:** 379–384

Friedman N (1998) The Bayesian structural EM algorithm. In *UAI*, Cooper G and Moral S (eds) pp 129–138. San Francisco, CA, USA: Morgan Kaufmann

Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen L, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier M-A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A-M, Schelder M, Schirle M, Remor M *et al* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440:** 631–636

Goldstein AL, McCusker JH (1999) Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. *Yeast* **15:** 1541–1553

Guarente L (1993) Synthetic enhancement in gene interaction: a genetic tool comes of age. *Trends Genet* **9:** 362–366

Han GS, Wu WI, Carman GM (2006) The Saccharomyces cerevisiae Lipin homolog is a Mg2 + -dependent phosphatidate phosphatase enzyme. *J Biol Chem* **281:** 9210–9218

Hartman JL, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. *Science* **291:** 1001–1004

Helenius A, Aebi M (2004) Roles of N-linked glycans in the endoplasmic reticulum. *Ann Rev Biochem* **73:** 1019–1049

Hillenmeyer M, Fung E, Wildenhain J, Pierce S, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, Altman R, Davis R, Nislow C, Giaever G (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320:** 362–365

Jonikas MC, Collins SR, Denic V, Oh E, Quan EM, Schmid V, Schwappach B, Walter P, Weissman JS, Schuldiner M (2009) Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science* **323:** 1693–1697

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucl Acids Res* **28:** 27–30

Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23:** 561–566

Kim W, Spear E, Ng D (2005) Yos9p detects and targets misfolded glycoproteins for ER-associated degradation. *Mol Cell* **19:** 753–764

Krogan N, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis A, Punna T, PeregrÃ–n-Alvarez J, Shales M, Zhang X, Davey M, Robinson M, Paccanaro A, Bray J, Sheung A, Beattie B *et al* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440:** 637–643

Moffat J, Grueneberg D, Yang X, Kim S, Kloepfer A, Hinkle G, Piqani B, Eisenhaure T, Luo B, Grenier J, Carpenter A, Foo SY, Stewart S, Stockwell B, Hacochen N, Hahn W, Lander E, Sabatini D, Root D (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124:** 1283–1298

Nakatsukasa K, Huyer G, Michaelis S, Brodsky JL (2008) Dissecting the ER-associated degradation of a misfolded polytopic membrane protein. *Cell* **132:** 101–112

Neal R (1998) Annealed importance sampling. *Stat Comput* **11:** 125–139

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc

Phillips PC, Otto SP, Whitlock MC, Wolf JD, Brodie EDI, Wade MJ (2000) Beyond the average: the evolutionary importance of gene interactions and variability of epistatic effects. In *Epistasis and the Evolutionary Process*, Wolf J, Brodie E, Wade M (eds) pp 20–38. Oxford, England: Oxford University Press

Qi Y, Suhail Y, Lin Y-Y, Boeke J, Bader J (2008) Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res* **18:** 1991–2004

Quan E, Kamiya Y, Kamiya D, Denic V, Weibezahn J, Kato K, Weissman J (2008) Defining the Glycan destruction signal for endoplasmic reticulum-associated degradation. *Mol Cell* **32:** 870–877

Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins S, Qu H, Shales M, Park H-O, Hayles J, Hoe K-L, Kim D-U, Ideker T, Grewal S, Weissman J, Krogan N (2008) Conservation and rewiring of functional modules revealed by an epistasis Map in fission yeast. *Science* **322:** 405–410

Schuldiner M, Collins S, Thompson N, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt J (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123:** 507–519

Schuldiner M, Metz J, Schmid V, Denic V, Rakwalska M, Schmitt HD, Schwappach B, Weissman JS (2008) The GET complex mediates insertion of tail-anchored proteins into the ER membrane. *Cell* **134:** 634–645

Segre D, Deluna A, Church G, Kishony R (2004) Modular epistasis in yeast metabolism. *Nat Genet* **37:** 77

Sprites P, Glymour C, Scheines R (1993) *Causation, Prediction, and Search*. New York: Springer-Verlag

St Onge R, Mani R, Oh J, Proctor M, Fung E, Davis R, Nislow C, Roth F, Giaever G (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39:** 199–206

Stefanovic S, Hegde R (2007) Identification of a targeting factor for posttranslational membrane protein insertion into the ER. *Cell* **128:** 1147–1159

Szathmary R, Bielmann R, Nitalazar M, Burda P, Jakob C (2005) Yos9 protein is essential for degradation of misfolded glycoproteins and may function as Lectin in ERAD. *Mol Cell* **19:** 765–775

Tong A, Lesage G, Bader G, Ding H, Xu H, Xin X, Young J, Berriz G, Brost R, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg D, Haynes J, Humphries C, He G, Hussein S, Ke L *et al* (2004) Global Mapping of the Yeast Genetic Interaction Network. *Science* **303:** 808–813

Typas A, Nichols R, Siegele D, Shales M, Collins S, Lim B, Braberg H, Yamamoto N, Takeuchi R, Wanner B, Mori H, Weissman J, Krogan N, Gross C (2008) High-throughput, quantitative analyses of genetic interactions in E. coli. *Nat Methods* **5:** 781–787

Williams C, Rasmussen C (1996) Gaussian processes for regression. In *Advances in Neural Information Processing Systems 8*, Touretzky D, Mozer M, Hasselmo M (eds) pp 514–520. Cambridge, Massachusetts, USA: MIT Press

Zupan B, Demsar J, Bratko I, Juvan P, Halter J, Kuspa A, Shaulsky G (2003) GenePath: a system for automated construction of genetic networks from mutant data. *Bioinformatics* **19:** 383–389